Optimizing Crop Yield Forecasting Using Hybrid Machine Learning Models

Narayanan Subbiah
Professor, Department of Information Technology SRM Valliammai Engineering College, Affiliated to Anna University, Chennai, Tamil Nadu, India.

***Abstract*—Crop yield prediction using machine learning enhances decision-making in agriculture by providing accurate and timely forecasts. This study utilized datasets including climate data, soil properties, satellite imagery, and historical yields. Four models were evaluated: Random Forest, XGBoost, Decision Tree, and Linear Regression. Random Forest Regression achieved the best performance with an R^2 of 0.97, MAE of 5,412, and RMSE of 12,450. XGBoost followed with an R^2 of 0.87, showing potential for improvement through hyperparameter tuning. Decision Tree Regression showed overfitting, with perfect training R^2 but a slightly lower test R^2 of 0.95. Linear Regression underperformed with an R^2 of 0.67, failing to capture complex patterns. Remote sensing and advanced analytics enhanced prediction accuracy and real-time monitoring. All models faced issues with infinite MAPE due to zero-yield values, highlighting the need for data cleaning. Random Forest proved to be the most reliable model, promoting efficient and sustainable agricultural practices.**

***Index Terms*—Machine learning, XGBoost,Mean Absolute Percentage Error, Regression**

I. INTRODUCTION

The global population is projected to grow from 7.9 billion in 2023 to 9.7 billion by 2050, increasing food demand by 44%. Rising incomes will shift diets toward more calorie- and protein-rich foods, especially meats and dairy, intensifying pressure on agriculture.

While technology offers promising solutions, challenges such as climate change, limited resources, and food insecurity persist. There's a growing push toward sustainable practices and healthier food systems, but faster progress is needed.

Accurate crop yield prediction is critical for efficient food production and economic stability, especially in the face of climate uncertainty. It enables informed decision-making by farmers and policymakers, helping manage resources and reduce waste.

Machine learning, remote sensing, and crop growth models improve prediction accuracy by analysing large, complex datasets including weather, soil, and satellite imagery. Deep learning, in particular, uncovers patterns that traditional models may miss, enhancing forecasting precision.

These technologies support real-time monitoring and early warnings for pests, weather events, and soil issues, reducing risks and enabling timely interventions. Yield prediction also aids in resource optimization—such as irrigation, fertilizers, and pesticides—promoting sustainability.

From a market perspective, accurate forecasts help balance supply and demand, reduce food waste, and stabilize prices. They also support financial planning, insurance, and investment decisions in agriculture.

At a policy level, governments can use predictive data for targeted subsidies, disaster relief, food security planning, and infrastructure development.

In conclusion, crop yield prediction is a transformative tool for improving agricultural efficiency, economic viability, and environmental sustainability. It supports a resilient food system by leveraging data and AI technologies.

II. RELATED WORK

Håkon Måløy a, et al in the research "Multimodal performers for genomic selection and crop yield prediction" introduces a promising deep learning framework for crop yield prediction, combining genomic data and weather patterns through the innovative Performer architecture. The results show that the Performer-based model significantly improves prediction accuracy over traditional methods, which can have a major impact on agricultural research, crop

breeding, and potentially other areas like animal breeding. The use of self-attention further enhances the model's interpretability, making it not only effective but also insightful for decision-making in breeding programs. [2]

Leelavathi Kandasamy Subramaniam a, *, Rajasenathipathi Marimuthu b "Crop yield prediction using effective deep learning and dimensionality reduction approaches for Indian regional crops"the paper proposes an advanced method for crop yield prediction (CYP) in southern India using deep learning (DL) and dimensionality reduction (DR) techniques. The approach is divided into three phases: preprocessing, where the agricultural data is cleaned and normalized; dimensionality reduction using Squared Exponential Kernel-based Principal Component Analysis (SEKPCA) to reduce data complexity; and crop yield prediction through a weight-tuned deep convolutional neural network (WTDCNN). The proposed method achieves an impressive accuracy of 98.96%, outperforming existing models. Its novelty lies in combining DL, DR, and WTDCNN to provide more precise and efficient predictions, benefiting agricultural planning and supporting improved farmer incomes [3]

Saeed Khaki1*, Lizhi Wang1 and Sotirios V. Archontoulis "A CNN-RNN Framework for Crop Yield Prediction Preprocessing "Crop yield prediction is a complex task due to its reliance on various factors such as crop genotype, environmental conditions, and management practices. This paper introduces a deep learning framework combining convolutional neural networks (CNNs) and recurrent neural networks (RNNs) to predict crop yields based on environmental and management data. The CNN-RNN model was tested alongside other popular methods, including random forest (RF), deep fully connected neural networks (DFNN), and LASSO, to forecast corn and soybean yields across the U.S. Corn Belt for the years 2016, 2017, and 2018. The CNN-RNN model significantly outperformed all other methods, achieving a root-mean-square-error (RMSE) of 9% and 8% of the respective average yields. Key features of the model include its ability to capture time dependencies of environmental factors and genetic improvements of seeds, generalize predictions to untested environments, and offer insights into how weather conditions, soil quality, and management practices impact crop yields. These advantages make the CNN-RNN model a promising tool for accurate crop yield prediction and agricultural decision-making.[5]

Priti Prakash Jorvekar1*, Sharmila Kishor Wagh2, Jayashree Rajesh Prasad3 "Predictive modeling of crop yields: a comparative analysis of regression techniques for agricultural yield prediction "this paper presents a comparative study on the performance of various regression models for crop yield prediction using a comprehensive dataset that includes historical crop yields, weather parameters, and pesticide data. The study evaluated multiple models, including Linear Regression, K-Nearest Neighbor Regression, Support Vector Regression, Decision Tree Regression, Random Forest Regression, Gradient Boosting Regression, and others, based on performance metrics such as R^2 score, RMSE, and computational time. The results showed that Random Forest Regression performed the best in terms of R^2, followed by K-Nearest Neighbor and Decision Tree Regression. However, the choice of the most suitable model also depends on factors like interpretability and computational efficiency. The findings provide valuable insights for farmers, policymakers, and researchers in selecting appropriate regression models for crop yield prediction. The study also suggests exploring the combination of regression models or integrating other machine learning techniques for improved prediction accuracy.[6]

Anikó Nyéki * and Miklós Neményi "Crop Yield Prediction in Precision Agriculture" predicting crop yields is a complex yet essential task in agriculture, influencing decision-making at various levels. It involves using factors like soil conditions, weather data, environmental influences, and crop parameters. Precision agriculture technologies, including sensors, management systems, and variable rate technologies, play a key role in enhancing crop yield and quality while minimizing environmental impact. Simulating crop yield helps understand the effects of deficiencies, pests, and diseases during the growing season. The integration of IoT and big data allows for more accurate yield predictions, with artificial intelligence further improving forecasting capabilities by analyzing vast amounts of agricultural data over time and space.[8]

Yuhan Wang 1,2, Qian Zhang 2, Feng Yu 2, *, Na Zhang 1,3, Xining Zhang 2, Yuchen Li 1, Ming Wang

2 and Jinmeng Zhang 2 “Progress in Research on Deep Learning-Based Crop Yield Prediction” crop yield prediction has become a key area of research in agricultural science, playing a vital role in economic development and policy formulation. Accurate predictions are essential for understanding the impact of factors like crop growth cycles, soil changes, and rainfall distribution. Traditional machine learning methods, while useful, often lack accuracy and show significant deviations from actual yields. This paper reviews the development of crop yield prediction, focusing on deep learning approaches. It analyzes various prediction models, their application to different crops, and offers suggestions for improving deep learning methods to enhance crop yield prediction in the future.[9]

III. SYSTEM DESIGN AND METHODOLOGY FOR CROP YIELD PREDICTION

Designing an effective crop yield prediction system involves a multi-layered architecture integrating data acquisition, preprocessing, machine learning model development, and deployment. The system begins with a Data Collection Module that aggregates information from diverse sources, including weather APIs, satellite imagery, soil sensors, and historical agricultural records. This is followed by a Data Preprocessing Module that cleans the dataset, handles missing values, encodes categorical variables, and scales numerical features using techniques such as Min-Max or Standard Scaling. The Feature Selection and Engineering Module identifies key variables—such as temperature, rainfall, soil nutrients, and crop types—and may generate additional features like Growing Degree Days (GDD) or Soil Moisture Index. The core of the system is the Machine Learning Module, which trains models such as Linear Regression, Random Forest, Support Vector Machines (SVM), XGBoost, or Long Short-Term Memory (LSTM) networks, depending on data characteristics. Model Evaluation and Optimization follows, using metrics like RMSE, MAE, and R^2, alongside hyperparameter tuning and cross-validation to improve performance. Once trained, the Prediction Module delivers yield forecasts based on input parameters, and results are presented through Visualization and Reporting tools such as interactive dashboards and plots to aid interpretation by farmers and policymakers. The system is deployed using platforms like Flask, Streamlit, or cloud services (AWS, Azure, GCP) for real-time, user-accessible predictions.

The system workflow begins with collecting environmental and agronomic data—such as average rainfall, temperature, and pesticide usage—followed by data cleaning and transformation. Feature engineering is used to extract meaningful patterns, which are then fed into various machine learning algorithms for training and evaluation. The best-performing model is selected based on cross-validation and error metrics. Predictions are generated and visualized, providing actionable insights for stakeholders. The dataset used includes 28,242 records spanning from 1990 to 2013, with five main variables: year, crop yield (hg/ha), average annual rainfall (mm), pesticide usage (tonnes), and average temperature (°C). The average yield is 77,053.33 hg/ha, but with a high standard deviation of 84,956.61 hg/ha, indicating substantial variability across regions and time periods. Rainfall averages 1,149.06 mm per year (SD: 709.81 mm), while pesticide use averages 1,149.06 tonnes with an extremely high SD of 59,958.78 tonnes, suggesting the presence of significant outliers. Average temperature is 20.54°C with a range from 1.3°C to 30.65°C. Due to the dataset’s wide variance and extreme values, preprocessing steps like outlier detection and feature scaling are critical for ensuring balanced model input and robust predictive performance.

Overall, this system design emphasizes the integration of diverse data sources, robust data processing, and advanced modeling techniques to deliver accurate and scalable crop yield predictions that support precision agriculture and policy planning.

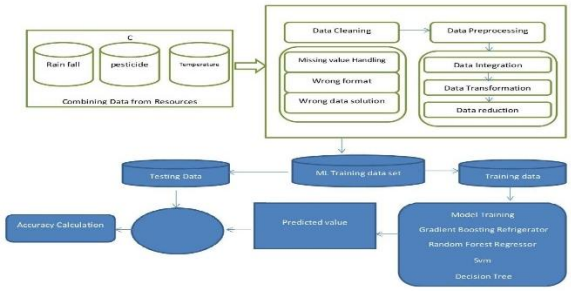


Figure:1

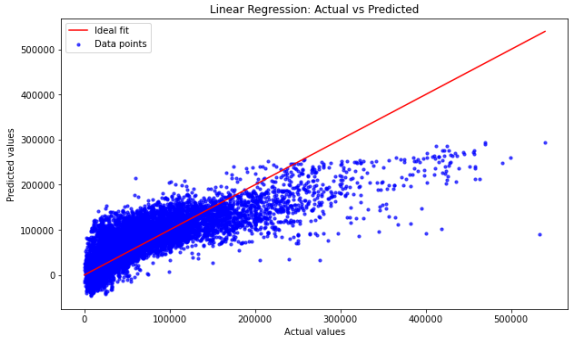
The dataset used for crop yield prediction contains 56,717 records and 12 columns, including essential attributes such as crop type (Item), region (Area), year, and yield (hg/ha_yield). To enhance clarity, the original Value column was renamed to hg/ha_yield, accurately reflecting crop yield in hectograms per hectare. Unnecessary columns such as codes and metadata were removed, leaving a cleaner dataset focused on the key variables for trend analysis. Yield values ranged from 0 to 1,000,000 hg/ha, with an average of over 62,000 and a high standard deviation, suggesting the presence of both outliers and missing data. Zero values could indicate missing entries, while extremely high values may reflect either outliers or exceptionally productive regions.

A refined dataset with 28,242 entries also included environmental features like annual rainfall, pesticide use, and average temperature, spanning the years 1990 to 2013. The crop yield in this subset showed even greater variability, averaging 77,053.33 hg/ha, ranging from just 50 to over 500,000 hg/ha. Rainfall ranged widely from 51 mm to 3,240 mm annually, while pesticide usage varied from 0.04 to 367,778 tonnes, indicating potential outliers requiring further inspection. Average temperatures ranged from 1.3°C to 30.65°C, reflecting diverse climatic zones. Given this variability, data scaling is essential to ensure balanced input to machine learning models.

Good data quality, including completeness and accurate pairing of environmental variables with yield, is crucial for reliable modeling. Several machine learning models are used for crop yield prediction, including Linear Regression, DecisionTreeRegressor, RandomForestRegressor, and XGBRegressor. Linear Regression, while simple and interpretable, struggles with complex, non-linear data and outliers. Decision Trees are flexible but prone to overfitting, whereas Random Forests combine multiple trees to reduce variance and improve generalization. XGBRegressor offers scalability and robustness, making it suitable for large, diverse datasets. Model performance is assessed using evaluation metrics such as MAE, MSE, RMSE, R^2, EVS, and MAPE, which help measure prediction accuracy, variance explained, and sensitivity to outliers. These metrics guide model selection and optimization for achieving better generalization and predictive accuracy in crop yield forecasting.

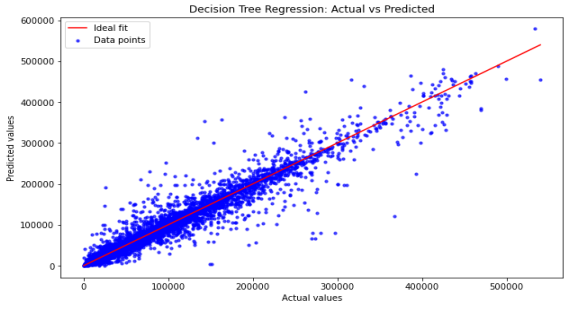
IV. RESULT ANALYSIS

linear regression model shows decent generalization, with an R^2 of ~0.67 for both training and test sets, meaning it explains 67% of variance in the target variable. However, the high MAE (~26,500) and RMSE (~39,000) indicate significant absolute prediction errors. The MAPE is infinite, likely due to division by zero, which can be addressed using SMAPE or filtering zero values. While the model isn't overfitting, performance could be improved through feature engineering, target transformation (e.g., log), trying advanced models (Random Forest, Gradient Boosting), or handling outliers.



Graph :1

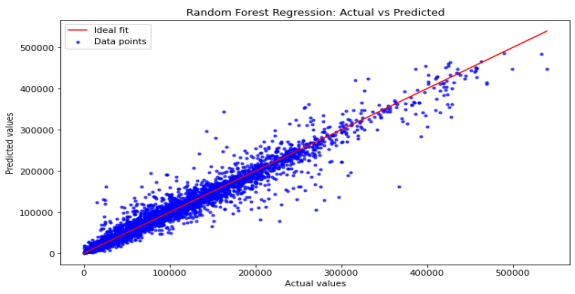
Decision Tree Regression model exhibits perfect performance on the training set ($R^2 = 1$, errors = 0), indicating overfitting. While the test set performance is much better than linear regression ($R^2 = 0.95$, RMSE = 15,428, MAE = 6,349), the large gap between training and test metrics confirms overfitting. The MAPE is still infinite, likely due to zero values in the target variable. To improve generalization, consider pruning the tree, setting a max depth, or trying ensemble methods like Random Forest or Gradient Boosting to balance bias and variance.



Graph :2

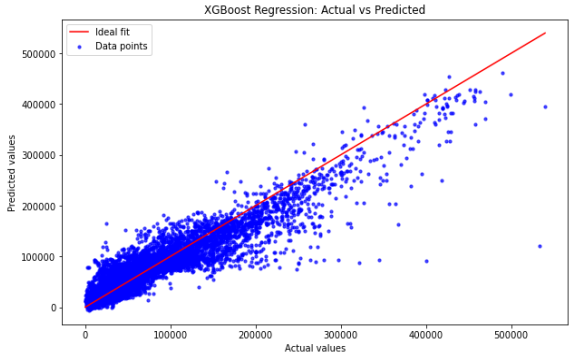
Random Forest Regression model significantly reduces overfitting compared to the Decision Tree, with $R^2 = 0.99$ on training and 0.97 on test, showing strong predictive power. The MAE (5,412) and RMSE (12,450) on the test set indicate much lower errors than both Linear Regression and Decision Tree. However, the MAPE remains infinite, likely due to zero values in the target variable. To further optimize, consider reducing tree depth, increasing estimators, or tuning hyperparameters to improve generalization while maintaining accuracy. Among the four models, Random Forest Regression performs the best, achieving the highest test R^2 (0.97) and the lowest test MAE (5,412) and RMSE (12,450), making it the most lanced in terms of accuracy and generalization.

Metric	Linear Regression	Decision Tree	Random Forest	XGBoost
Training R^2	0.6667	1.0000 (Overfit)	0.9949	0.8786
Test R^2	0.6702	0.9485	0.9664	0.8722
Training RMSE	39,148	0 (Overfit)	4,834	23,627
Test RMSE	39,021	15,428	12,450	24,288
Training MAE	26,660	0 (Overfit)	1,953	14,740
Test MAE	26,524	6,349	5,412	15,181
MAPE	inf	inf	inf	inf



Graph :4

XGBoost Regression model delivers a significant improvement over Linear Regression, achieving $R^2 = 0.88$ (train) and 0.87 (test), meaning it explains about 87% of variance while maintaining good generalization. The MAE (~15,000) and RMSE (~24,000) are lower than Linear Regression but higher than Random Forest, suggesting room for further optimization. The MAPE remains infinite, likely due to zero values in the target variable. To further refine performance, consider hyperparameter tuning (learning rate, max depth, boosting rounds), feature selection, or handling outliers.



Graph :3

XGBoost follows closely with an R^2 of 0.87 but slightly higher errors, indicating room for improvement through hyperparameter tuning. Decision Tree Regression completely overfits ($R^2 = 1$ on training) but still generalizes well ($R^2 = 0.95$ on test), though it may be unstable without pruning. Linear Regression performs the worst ($R^2 = 0.67$) with significantly higher errors, suggesting it struggles to capture complex relationships in the data. A common issue across all models is infinite MAPE, likely due to zero values in the target variable, which should be addressed separately. Overall, Random Forest is the most reliable choice, but XGBoost could be further optimized for better performance

R^2 Score Comparison: Higher is better. Random Forest and Decision Tree perform best on the test set, while Linear Regression is the weakest.

MAE Comparison: Lower is better. Random Forest has the lowest test MAE, meaning it makes the least absolute errors.

RMSE Comparison: Lower is better. Again, Random Forest has the lowest RMSE, indicating better generalization.

From these visualizations, Random Forest is the best overall model, while Linear Regression struggles with high errors.

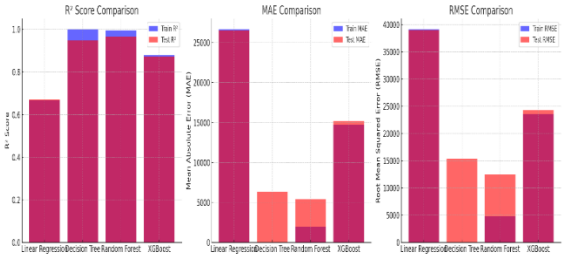


Figure:2

REFERENCES

[1] Thomas van Klompenburga, Ayalew Kassahuna, Cagatay Catalb "Crop yield prediction using machine learning: A systematic literature review" Computers and Electronics in Agriculture, Volume 177, October 2020, 105709.

[2] Håkon Måløy a, ∗, Susanne Windju b, Stein Bergersen b, Muath Alsheikh b, c, Keith L. Downing "Multimodal performers for genomic selection and crop yield prediction"Smart Agricultural Technology 1 (2021) 100017.

[3] Leelavathi Kandasamy Subramaniam a, *, Rajasenathipathi Marimuthu "Crop yield prediction using effective deep learning and dimensionality reduction approaches for Indian regional crops"e-Prime - Advances in Electrical Engineering, Electronics and Energy 8 (2024) 100611

[4] R. Ghadge, J. Kulkarni, P. More, S. Nene and R. L. Priya, "Prediction of crop yield using machine learning", Int. Res. J. Eng. Technolgy, vol. 5, 2018.

[5] F. H. Tseng, H. H. Cho and H. T. Wu, "Applying big data for intelligent agriculture-based crop selection analysis", IEEE Access, vol. 7, pp. 116965-116974, 2019

[6] M. Alagurajan and C. Vijayakumaran, "ML Methods for Crop Yield Prediction and Estimation: An Exploration", International Journal of Engineering and Advanced Technology, vol. 9, no. 3, 2020

[7] K. A. Shastry and H. A. Sanjay, "Hybrid prediction strategy to predict agricultural information", Applied Soft Computing, vol. 98, pp. 106811, 2021.

[8] T. Senthil Kumar, "Data Mining Based Marketing Decision Support System Using Hybrid Machine Learning Algorithm", Journal of Artificial Intelligence, vol. 2, no. 03, pp. 185-193, 2020.